Tundrai nyenyec morfológiai elemző és generátor

Novák Attila[1] és Wenszky Nóra[2]

MorphoLogic Kft. 1126 Budapest Orbánhegyi út 5.
[1]novak@morphologic.hu,
[2]nora@nytud.hu

Kivonat: Ebben a cikkben egy az Uráli nyelvcsalád északi szamojéd ágához tartozó *tundrai nyenyec* nyelvű szóalaktani elemzőprogram és szóalakgenerátor létrehozásáról számolunk be. A nyelv bemutatása és az elemző alapjául szolgáló korábbi munkák ismertetése után részletesen tárgyaljuk az elemzőprogram egyes moduljait és működését. Az elemző lexikona mintegy 19 500 tő mögöttes fonológiai reprezentációját tartalmazza, melyek 266 különböző ragozási osztályba sorolhatók. A toldaléktár 254 mögöttes toldalékalakot tartalmaz. Az elemző a nyelv inflexiós jelenségeit teljes körűen kezeli, beleértve az általában a szóképzés körében tárgyalt igenevek és gerundiumok kezelését is.

1 Bevezetés

A tundrai nyenyec szóalaktani elemzőprogram, mely erre a nyelvre az első ilyen eszköz, egy olyan projektum[65] részeként valósult meg, melynek célja korpuszok, morfológiai elemzőprogramok és egyéb elektronikus nyelvi erőforrások létrehozása volt néhány kisebb, az uráli nyelvcsaládba tartozó nyelven. A projektum keretében a tundrai nyenyec mellett a nganaszan, a komi, az udmurt, a mari és a manysi nyelvekre készült morfológiai elemzőprogram (Novák, 2004 [2]; Prószéky és Novák, 2005 [3]).

2 A tundrai nyenyec nyelv

A projektum keretében leírt másik északi szamojéd nyelvet, a nganaszant, közvetlen kihalás fenyegeti, beszélőinek száma már csak mintegy 500 fő. A tundrai nyenyecnek ezzel szemben mintegy 25 000 beszélője van. Ugyanakkor ez a beszélőközösség hatalmas területen oszlik el. A tundrai nyenyecek hagyományos lakóterületét nyugaton a Kanyin-félsziget, keleten a Jenyiszej deltája határolja. A XIX. és a XX. század-

[65] Komplex Uráli nyelvészeti adatbázis, NKFP 5/135/2001. A projektumban a Nyelvtudományi Intézet Finnugor Osztálya, különböző finnugor nyelvészeti tanszékek és a MorphoLogic Kft. vett részt.

ban Novaja Zemljára, ill. nyugaton még a Kola-félszigetre is betelepültek, keleten pedig a Tajmir-félszigeten is megjelentek. Míg keleten a tundrai nyenyec nyelv jelenleg is expanzióban van (más szamojéd nyelvek rovására), keleten jellemző a nyelvvesztés a tundrai nyenyecek körében. A nagy földrajzi távolságok ellenére a tundrai nyenyec dialektálisan nem nagyon tagolt: a beszélők mind viszonylag könnyen megértik egymást. A nyelv három nagy dialektuscsoportra (nyugati, középső és keleti) oszlik. A nyugati dialektusok jobban különböznek a középső és keleti változatoktól, mint ezek egymástól.

3 A források

Az elemző készítésénél a *Tapani Salminen* által használt latin betűs fonologikus átírást használtuk, és az ő leírására támaszkodtunk a morfofonológiai szabályok megfogalmazásánál is (Salminen, 1997 [4] és 1999 [6]). Salminen elsősorban a középső dialektuscsoport nyelvjárását írja le (bár külön kitér a nyugati változat jellegzetességeire is), így a mi elemzőnk is erre a változatra készült. Rendelkezésünkre bocsátotta morfológiai szótára anyagát (Salminen, 1998=MDTN [5]) és elküldte disszertációja szövegét is (Salminen, 1997=TNI [4]) számítógéppel kereshető formában, ami felbecsülhetetlen segítséget jelentett az elemző elkészítésében, mert a számítógéppel gyorsan meg lehetett találni a szövegben egyébként meglehetősen szétszórva előforduló információkat.

A tundrai nyenyec – és általában az északi szamojéd nyelvek – esetében a nyelvészeti leírás szempontjából talán a legnagyobb problémát a rendkívül bonyolult és produktív felszíni fonológiai–fonetikai folyamatok jelentik. Ezek nemcsak a formális számítógépes modell megalkotása és implementálása szempontjából, hanem már a nyelvi adatok puszta lejegyzése és bármilyen elfogadható grammatikai modell megalkotása szempontjából is nehézséget jelentenek.

Ami ezekben a nyelvekben egy morféma összes megjelenési formájában közös (a morféma „mögöttes reprezentációja"), az gyakran valami annyira absztrakt dolog, hogy első ránézésre szinte semmi köze nincs a morféma konkrét megjelenési formáihoz, az allomorfjaihoz. Ennek persze az az oka, hogy maguk az allomorfok sem hasonlítanak első ránézésre egymáshoz. Ezért ezeknek a nyelveknek a fonológiájáról és morfológiájáról rendkívül nehéz volt adekvát leírást készíteni, és csak a legutóbbi időben születtek meg ezek a modellek. Salminen leírására azért esett a választásunk, mert ezt a bonyolult absztrakciós folyamatot következetesen végigvitte, és konzisztensnek tűnő, ugyanakkor elég jól formalizált leírást alkotott a tundrai nyenyec morfológiáról. Salminen leírásának magas színvonala és formalizáltságának foka ritka az uráli nyelvészet körébe tartozó nyelvleírások között.

3.1 Salminen jelölésmódja

Két fonéma mind a nganaszanban, mind a tundrai nyenyecben rendkívül problematikus: a schwa (Salminen terminológiájában a tundrai nyenyec esetében „redukált magánhangzó") és a gégezárhang. Az előbbi fonetikailag rendkívül instabil, és a felszínen gyakran a magánhangzókészlet más elemeihez hasonul (illetve a tundrai nyenyecben ezen felül meglehetősen bonyolult szabályoknak engedelmeskedve tűnik el

vagy jelenik meg), és általában csak az egész nyelv fonológiai és morfológiai rendszerének figyelembevételével lehet egy-egy konkrét esetben a kilétét megállapítani. A tundrai nyenyecben ráadásul az *a* fonéma is igen változékony fonetikailag (a hossza és a minősége szempontjából) a hangsúlyviszonyok függvényében. A nyenyecben emellett legalább két különböző gégezárhang van, amelyek a fonetikai realizáció szempontjából nem különböznek egymástól, de különböző fonológiai környezetekben különbözőképpen viselkednek. A „nazalizálható gégezárhang" csak szünet előtt jelenik meg a felszínen gégezárként, egyébként obstruensek előtt homorgán nazálisként realizálódik, szonoránsok előtt pedig eltűnik. A „nem nazalizálható gégezárhang" ezzel szemben az obstruensek előtt tűnik el. Van a gégezárhangnak egy harmadik típusa is: ez nem egy fonológiailag jelen lévő szegmentum felszíni megvalósulása, hanem egy automatikus felszíni fonetikai folyamat eredményeként jelenik meg: a szünet előtti szóvégi mássalhangzó fonémák után toldódik be (hogy ez akusztikailag hogyan realizálódik, arra vonatkozólag nem állnak rendelkezésünkre adatok). A gégezárhanggal kapcsolatban még egy érdekesség megjegyzendő: intervokálisan csupán néhány szóban jelenik meg.

Salminen arra törekedett, hogy a nyelv toldalékolási paradigmáinak figyelembevételével olyan jelölésmódot hozzon létre, amely konzisztens, összhangban van a szavak fonemikus felépítésével, és tükrözi a paradigmatikus viselkedésüket is. Például a fonológiailag schwára (Salminen terminológiájában „redukált magánhangzóra") végződő szavak végén a magánhangzó fonetikailag általában nem realizálódik, de ettől a fonetikai ténytől eltekintve ezek a szavak fonológiailag és morfológiailag (a ragozási paradigmájuk szempontjából) teljes mértékben magánhangzó végű szavakként viselkednek, ezért Salminen ilyenként is ábrázolja őket. Ráadásul ezekben az esetekben az automatikus szóvégi gégezárhang-betoldás is mindig elmarad.

A cirill betűs tundrai nyenyec helyesírás a Salminen által használt fonologikus jelölésrendszerrel szemben erősen fonetikus, a hangsúlytalan a-k és schwák hasonulását az előző magánhangzókhoz általában jelöli, csakúgy, mint a betoldott fonetikai schwákat és gégezárhangokat, ugyanakkor mindössze öt magánhangzót különböztet meg. Mindezt Salminen szerint meglehetősen inkonzisztens módon. Ugyanakkor a helyesírás egyáltalán nem jelöli a mennyiségi (hosszúságbeli) különbségeket, azokat sem, amelyek tényleges fonológiai különbségeken alapulnak. Például a fonológiailag ténylegesen gégezárhangra végződő szavak esetében a gégezárhangot megelőző mássalhangzó akusztikailag jelentősen hosszabb, mint a szóvégi mássalhangzók után automatikusan betoldott gégezárhangok előtt, de ez az írásban nem tükröződik.

Mivel Salminen meglehetősen absztrakt jelölésmódjában a szavak „felszíni alakja" is csak nagyon távoli, áttételes és távolról sem egyértelmű viszonyban van ugyanezen szavak ortográfiai alakjával, ezért egyrészt az oroszországi nyelvészek kétkedéssel tekintenek Salminen jelölésmódjára, másrészt a mi elemzőnket sem lehet egyelőre közvetlenül írott nyenyec szövegek elemzésére használni. Annak természetesen nincs elvi akadálya, hogy az elemző szabályrendszerét kiegészítsük azokkal a szabályokkal, amelyek a Salminen-féle felszíni reprezentáció és a szavak ortográfiai alakjai közötti űrt áthidalják, de ez nem egészen triviális feladat, és egyelőre nem képezi részét leírásunknak.

4. A morfológiai elemző

Az elemzőt a Xerox cég reguláris relációkalkuluson alapuló morfológiai fejlesztőrendszerének, az *xfst*-nek (Xerox Finite-State Tool) felhasználásával készítettük el (Beesley–Karttunen, 2003 [1]). Ez a generatív fonológusok által megszokott kontextusfüggő újraírószabály-formalizmussal leírt szekvenciális fonológiai szabályegyüttesek megadását teszi lehetővé, és kiszámítja az egyes szabályok egymással, illetve a lexikonnal való komponálásával előálló teljes morfofonológiai leírást egyetlen kétszintű véges állapotú fordítóautomata formájában.

Miután a nganaszan elemző és szólak-generátor elkészítéséhez is ezt az eszközt használtuk (Novák, 2004 [2]), logikus döntésnek tűnt, hogy a nganaszannal közeli rokonságban álló, és hasonló bonyolultságú tundrai nyenyec esetében is ehhez a megoldáshoz folyamodjunk.

A morfológiai elemző tőadatbázisának alapjául az MDTN szótár [5] szolgált, a toldaléklexikon, a tő- és a toldaléktárat összekapcsoló általánosabb paradigmatípus-osztályozás és a fonológiai szabálykomponens pedig Salminen disszertációja (TNI=Salminen, 1997 [4]) és a *Grammatical sketch* (Salminen, 1999 [6]) alapján készült. A következőkben részletesen bemutatjuk az elemző moduljait: a tőtárat, a toldaléktárat és a szabályfájlt.

4.1 A tőtár

Az MDTN szótárt Microsoft Excel formátumban kaptuk meg. Az szótárt ISO-8859-1 kódolású szövegfájllá konvertáltuk. A konvertált szótár egy részlete alább látható:

```
   nga°                                Part          NGA ' Ø
@  (E -ibø- ~) W-C søb°bø-             Vt ¢ ø»yi     SØPØ 0«MPØ
@  (-mpø- ~ E -ibø- ~) W-C tyeb°bø-    Vt ¢ ø»yi     TYEPØ 0«MPØ
@  (E -ibø- ~) W-C tyíb°bø-            Vt ¢ ø»yi     TYÍPØ 0«MPØ
   (E -ibø- ~) W-C løbc°bø-            Vt ¢ ø»yi     LØPSØ 0«MPØ
@  (E -ibø- ~) W-C yabc°bø-            Vt ¢ ø»yi     JAPSØ ' 0«MPØ
@  (-mpø- ~ E -ibø- ~) W-C nyanc°bø-   Vt ¢ ø»yi     NYAHSØ 0«MPØ
@  (-mpø- ~) nyenc°bø-                 Vi ¢ ø»yi     NYEHSØ 0«MPØ
@  (E -ibø- ~) W-C syenc°bø-           Vt ¢ ø»yi     SYEHSØ '' 0«MPØ
   sødøb°                              N ø=          SØTØPØ
```

A tőtár létrehozásához a mögöttes alak mezőt (utolsó oszlop), a kategóriamezőt (Part, Vt ¢, N stb.) és a toldalékolási osztály (ø»yi, ø= stb.) mezőket használtuk. A felszíni alak mezőre nem volt szükségünk.

A mögöttes alak morfémákra van szegmentálva. A '-ok a homonim tövek azonosítására szolgának. A képzők illesztésénél fellépő sandhijelenségekre utaló jeleket (pl. 0«) a mögöttes alak tartalmazza a produktív hasonulási jelenségek kivételével. (A produktív hasonulásokat a külön leírt fonológiai és morfofonológiai szabálykomponens kezeli.)

A fenti adatbázisnak a Xerox rendszer által a lexikon leírására használt *lexc* formátumára való átalakításához készítettünk néhány programot. Az átalakító a fenti szótárrészletet először az alábbi formára konvertálja:

```
Root      nga^1|ø                          Part_
Root      søpø|^0«mpø                      Vt=c_ø»yi
Root      tyepø|^0«mpø                     Vt=c_ø»yi
Root      tyípø|^0«mpø                     Vt=c_ø»yi
Root      løpsø|^0«mpø                     Vt=c_ø»yi
Root      japsø^1|^0«mpø                   Vt=c_ø»yi
Root      nyahsø|^0«mpø                    Vt=c_ø»yi
Root      nyehsø|^0«mpø                    Vi=c_ø»yi
Root      syehsø^2|^0«mpø                  Vt=c_ø»yi
Root      søtøpø                           N_ø=
```

Az első oszlop azt adja meg, hogy az adott morfémasorozat melyik allexikonba kerül. Ezt követi a morfémasorozat mögöttes alakja (| jelekkel morfémákra szegmentálva). Végül az adott morfémasorozat folytatási osztálya áll (ez egy lexikonnév: minden olyan morfémasorozat, ami az adott nevű lexikonban szerepel, követheti az adott morfémasorozatot). A Root lexikonban szereplő elemek állhatnak a szó elején. A szó végét a # folytatási osztály jelzi. A következő lépésben az egyes allexikonokba tartozó elemeket egybegyűjtöttük a *lexc* formátumnak megfelelően. Az így kapott adatbázis lett az elemzőprogram tőtára.

4.2 A ragozási osztályok leírása

Salminen ragozási osztályait az egyes kategóriákon belül gyakorisági sorrendbe rendeztük és a TNI-ben felállított általánosabb toldalékolási osztályokba soroltuk. Az egyes osztályok jellegzetességeit az MDTN bevezetője elég részletesen tárgyalja. A besorolás alapja alapvetően ez a leírás volt. Ugyanakkor az MDTN bevezetője számos olyan tényt is közöl az egyes osztályoknál, amelyek jóval általánosabb fonológiai, ill. morfofonológiai folyamatok következményei. Ezeknek természetesen nem itt van a helyük, hanem a fonológiai, ill. morfofonológiai szabályok leírásánál. Itt ki kellett hagynunk ezeket a redundáns információkat. Az alábbi lista néhány tárgyasvisszaható és tárgyas igeosztály így annotált leírását mutatja, az osztály neve után annak lexikonbeli gyakorisága és szoknak az általánosabb toldalékolási osztályoknak a felsorolása következik, amelyekbe az adott toldalékolási osztály tartozik (részlet a teljes toldalékolásiosztály-listából):

```
Vt-r µ a: 162  POLYVSTEMV
Vt-r µ ye: 43  POLYVSTEMV
Vt-r µ l: 28    CSTEMV
Vt-r µ r: 15 CSTEMV
Vt-r µ ø: 13   POLYVSTEMV  ØSTEMV
...
Vt ¢ ø»yi: 2361   ALTV ALTVøi
Vt µ a: 1380   POLYVSTEMV
Vt µ ye: 579  POLYVSTEMV
```

Írtunk egy programot, amely ezt a leírást folytatási osztály alapú lexikonná alakítja. Ez a lexikonrészlet kapcsolja össze a tőlexikon folytatási osztályait az alább ismertetendő toldaléklexikon toldalékosztályaival. Az alábbi példa a leggyakoribb tárgyasvisszaható igeosztály és a leggyakoribb tárgyas igeosztály ilyen formájú leírását mutatja. A középső oszlopban []-ben szereplő címkéket az adott osztályba tartozó szavak elemzésekor az elemző kiírja. A @ jelek közötti jegyeket a program a tőhöz kapcsol-

ható toldalékosztályok szűrésére használja. Az alábbi példákban szereplő `@P.CONJ.tr@` szimbólum például a `CONJ` (igeragozás) jegy `t-r` (tárgyas-reflexív) értékre való beállítását írja elő. Ez lehetővé teszi valamennyi igei személyragosztálynak az ilyen osztályú tövekhez való kapcsolódását. Ugyanakkor pl. a `@P.CONJ.t@` (tárgyas) osztályú tövekhez a visszaható toldalékok nem kapcsolódhatnak.

```
#Class Vt-r=m_a - frq=162
Vt-r=m_a            [V][t-r][Mom]@P.CONJ.t-r@  Vt-r=m_a_
Vt-r=m_a_           POLYVSTEMV
Vt-r=m_a_           V_base
...

#Class Vt=c_ø»yi - frq=2361
Vt=c_ø»yi           [V][t][Cnt]@P.CONJ.t@      Vt=c_ø»yi_
Vt=c_ø»yi_          ALTV
Vt=c_ø»yi_          V_base
Vt=c_ø»yi_          »^yi          GFS=SFS
```

4.3 A toldaléktár

Toldaléklexikonunk és a szabályok leírásának alapjául a TNI és a *Grammatical sketch* szolgált. Ezeket a műveket többször elolvastuk, kijegyzeteltük, és több plakát méretű színes gráfot rajzoltunk, amelyek az egyes tőalakváltozatok (pl. az igéknél 'Special Finite Stem'[66] , 'General Finite Stem'[67], 'General and Special Modal Substems'[68]) és az egyes toldalékok alakváltozatait (itt nem figyelembe véve az általános fonológiai és morfofonológiai folyamatok következtében előálló alternációkat) és egymáshoz viszonyított sorrendjét ábrázolták. Mivel a nyenyec toldalékolási rendszer igen bonyolult, hasonlóan a nganaszanéhoz, ezek a gráfok tekintélyes méretűek lettek.

Mivel a forrásművek a képzőket nem tárgyalták, elemzőnk csak a ragozást kezeli produktívan. (Salminen a ragozás körébe utalta az igenevek és gerundiumok képzését, amelyeket általában képzett alakoknak tekintik, ezeket tehát elemzőnk kezeli). Ugyanakkor az MDTN szótár rengeteg képzett alakot tartalmaz (méghozzá morfémákra szegmentálva), ez tehát majd (egy következő projektum keretében) jó alapja lehet a produktív szóképzési folyamatok formális leírásának és számítógépes modellezésének.

A toldalékolási gráfokat ábrázoló plakátok alapján készítettük el a toldaléklexikont a *lexc* formátumához közeli fentebb ismertetett folytatási osztályokon alapuló formában. Ez lexikonunk harmadik része, melyet az előbb ismertetett két résszel (tőtár, toldalékolási osztályok tára) összefűzve a teljes lexikont megkaptuk. A toldalékleírás tartalmaz néhány javítást a forrásokban leírtakhoz képest, amelyeket az elemző írása, illetve tesztelése során talált hibák alapján Salminennel konzultálva végeztünk.

A toldaléktárban azoknak a jelenségeknek a kezelésére, amelyek nem szomszédos morfémák közötti megszorításokon alapulnak, jegyérték-ellenőrző kifejezéseket használtunk. Ilyenek gondoskodnak például arról, hogy a megfelelő igeiszemélyragosztályok éppen a megfelelő vonzatkeret-osztályba tartozó tövekhez járulhassanak

[66] SFS = a visszaható és a többes számú tárgyas kijelentő módú finit igealakok töve

[67] GFS = a többi finit igealak töve

[68] ugyanez a különböző egyéb igemódokra

csak (az alanyi ragozás személyragjai a reflexív igéket kivéve mindhez, a tárgyas személyragok a tárgyas és a tárgyas-reflexív igékhez, a visszaható személyragok pedig a reflexív és a tárgyas-reflexív igékhez). Hasonlóan kezeltük az enklitikus múltidő-jel megjelenésére vonatkozó megszorításokat, az opcionális palatalizáció és az *e~i°* alternáció jelenségét.

4.4 A szabályfájl

Salminen leírását a nyenyec morfofonológiáról viszonylag egyszerűen le lehetett fordítani az *xfst* szabályformalizmusára. Ugyanakkor a kézzel írott nyelvtanokban általában sok részlet homályban marad. Itt is ez volt a helyzet. Mindazokat a pontokat, ahol az eredeti leírás homályos volt (pl. a szabályok sorrendezése, alkalmazási köre, a környezet pontos leírása, egyáltalán formális megfogalmazása, a kivételek stb.) explicitté kellett tennünk. Az általunk program formájában implementált leírás a tundrai nyenyec fonológiáról és morfofonológiáról tehát az eredeti forrásokban közölt leírások jóval részletesebb és teljes formális igényű javított változatának tekinthető.

A szabályfájl definíciókkal (főleg szegmentumosztályok definícióival) kezdődik. Ezt követik a fonológiai, majd a morfofonológiai folyamatokat leíró szabályok. A fonológiai szabályok közül az igen komplex magánhangzó-redukciós folyamatot leíró szabály az első, ezt követik a különböző produktív mássalhangzó-hasonulási szabályok. A morfofonológiai szabályok az ingadozó szegmentumok viselkedését leíró szabályokkal kezdődnek, ezeket követik a különböző korlátozott morfológiai környezetekben működő (általában csak morfémahatáron, ill. csak egy-egy morféma környezetében működő) sandhiszabályok.

A szabályok sorrendezése a szabályok definíciójától függetlenül van megadva, hiszen ezt a fejlesztés–tesztelés során többször meg kellett változtatnunk.

4.5 A kész elemző

Ugyanarra a nyelvtanra alapozva több változatot is elkészítettünk az elemzőből, amelyek az elemzések „szószátyársága” szempontjából különböztek egymástól: a kevésbé bőbeszédű változat csak az abszolút tő szótári alakját és a tő és a toldalékok morfoszintaktikai jegyeit írja ki, a gazdagabb változat kiírja a toldalékok mögöttes alakját is. Tulajdonképpen az utóbbi az alapvető változat, az előbbit ebből egy megfelelő szűrő hozzákomponálásával állítjuk elő. A tömörebb elemzéseket adó változat inverzét használjuk szóalak-generátorként. Az interaktív tesztelés céljára készítettünk olyan változatokat is, amelyek a Salminen által használt karaktereknél a billentyűzeten könnyebben gépelhető (ASCII) karakterekkel dolgoznak, de egyébként ekvivalensek a Salminen jelölésrendszerét használó változattal.

Néhány példa az egyes változatok által adott elemzésekre:

ASCII tömör elemző

```
myeryuj'w@nantyih       myeryo|>^yuj'[N][Poss][Pros][Sg][3][Du]
myeryuj'w@nantyih       myeryo|>^yuj'[N][Poss][Pros][Sg][2][Du]
```

ASCII bőbeszédű elemző

```
myeryuj'w@nantyih
myeryo|>^yuj'[N][Poss]m'na[Pros][Sg]ht[3]yih[Du]
myeryuj'w@nantyih
myeryo|>^yuj'[N][Poss]m'na[Pros][Sg]ht[2]yih[Du]
```

ASCII generátor

```
myeryo|>^yuj'[N][Poss][Pros][Sg][3][Du] myeryuj'w@nantyih
```

Salminen tömör elemző

```
myeryuj°wønantyih       myeryo|»^yuj°[N][Poss][Pros][Sg][3][Du]
myeryuj°wønantyih       myeryo|»^yuj°[N][Poss][Pros][Sg][2][Du]
```

Salminen bőbeszédű elemző

```
myeryuj°wønantyih
myeryo|»^yuj°[N][Poss]m°na[Pros][Sg]ht[3]yih[Du]
myeryuj°wønantyih
myeryo|»^yuj°[N][Poss]m°na[Pros][Sg]ht[2]yih[Du]
```

5 Összefoglalás

Az elemző lexikona mintegy 19 500 tő mögöttes fonológiai reprezentációját tartalmazza. Ezeket Salminen 266 különböző ragozási osztályba sorolta. A toldaléktár 254 mögöttes toldalékalakot tartalmaz. Az elemző a nyelv inflexiós jelenségeit teljes körűen kezeli, beleértve az általában a szóképzés körében tárgyalt igenevek és gerundiumok kezelését is. A tőtárban szereplő rengeteg képzett szó morfémákra tagolását is megadja a program, annak ellenére, hogy a szóképzést produktívan nem kezeli, hiszen ez az információ az eredeti forrásban szerepelt, és a tőtár létrehozása során megtartottuk.

A létrehozott eszközök megfelelő nyelvi adatok megléte esetén a bennük implementált nyelvtan adekvátságának messzemenő tesztelését teszik lehetővé, olyan alapossággal, amely – különösen egy a tundrai nyenyechez hasonló bonyolultságú nyelv esetében – kézzel elképzelhetetlen. Az elemzőt elsőként az MDTN szótár előszavában megjelent teljes példaparadigmákon teszteltük. Az első változatban természetesen találtunk jó pár hibát, a lexikonban, a szabályok megfogalmazásában és ezek sorrendezésében is. Ezeket kijavítottuk. Másik tesztanyagként a TNI-ben szereplő példaszavakat tudtuk használni, hiszen ott is szerepel a szavak kézi elemzése is.

Sajnos nagyon kevés olyan korpusz áll rendelkezésre, amely Salminen jelölésmódjával van lejegyezve. Ezért is nagyon kívánatos lenne, hogy az elemzőt kiegészítsük azokkal a leképezési szabályokkal, amelyek lehetővé teszik, hogy közvetlenül a nyenyec ortográfiával lejegyzett szövegek elemzésére használhassuk. A másik lehetséges továbbfejlesztési lehetőség a képzők produktív kezelése, feltéve, hogy a szükséges ismeretek rendelkezésre állnak majd.

Salminennel egyébként a fejlesztés során végig szoros kapcsolatban voltunk és minden pontatlanságra és hiányra felhívtuk a figyelmét a forrásokban, amire az elem-

ző fejlesztése során fény derült. Reményeink szerint így együttműködésünk az ő számára is hasznos volt.

Bibliográfia

1. Beesley, Kenneth R. and Lauri Karttunen: *Finite State Morphology*, CSLI Publications, Ventura Hall (2003)
2. Novák Attila: Az első nganaszan szóalaktani elemző. In: *II. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2004).* Szegedi Tudományegyetem (2004) 195–202
3. Prószéky, Gábor and Attila Novák: Computational Morphologies for Small Uralic Languages. In: A. Arppe, L. Carlson, K. Lindén, J. Piitulainen, M. Suominen, M. Vainio, H. Westerlund, A. Yli-Jyrä (eds.): *Inquiries into Words, Constraints and Contexts. Festschrift in the Honour of Kimmo Koskenniemi on his 60th Birthday*. Gummerus Printing, Saarijärvi/CSLI Publications, Stanford (2005) 116–125
4. Salminen, Tapani: *Tundra Nenets inflection.* Mémoires de la Société Finno-Ougrienne 227; Helsinki (1997) = TNI
5. Salminen, Tapani: *A morphological dictionary of Tundra Nenets.* Lexica Societatis Fenno-Ugricae 26; Helsinki (1998) = MDTN
6. Salminen, Tapani: *Tundra Nenets (A grammatical sketch)*, http://www.helsinki.fi/~tasalmin/sketch.html (1999)